Neuromorphic Detection of Vowel Representation Spaces

Pedro Gómez-Vilda[1], José Manuel Ferrández-Vicente[2], Victoria Rodellar-Biarge[1], Agustín Álvarez-Marquina[1], Luis Miguel Mazaira-Fernández[1], Rafael Martínez-Olalla[1], and Cristina Muñoz-Mulas[1]

[1] Grupo de Informática Aplicada al Tratamiento de Señal e Imagen, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28660 Madrid
pedro@pino.datsi.fi.upm.es
[2] Dpto. Electrónica, Tecnología de Computadoras, Univ. Politécnica de Cartagena, 30202, Cartagena

Abstract. In this paper a layered architecture to spot and characterize vowel segments in running speech is presented. The detection process is based on neuromorphic principles, as is the use of Hebbian units in layers to implement lateral inhibition, band probability estimation and mutual exclusion. Results are presented showing how the association between the acoustic set of patterns and the phonologic set of symbols may be created. Possible applications of this methodology are to be found in speech event spotting, in the study of pathological voice and in speaker biometric characterization, among others.

1 Introduction

Speech processing is evolving from classical paradigms more or less statistically oriented to psycho- and physiologic paradigms more inspired in speech perception facts [1]. Especially important within speech perception are vowel representation spaces. These may be formally defined as applications between the space of acoustic representations at the cortical level to the set of perceptual symbols defined as vowels at the phonologic or linguistic level [12]. These relations can be expressed using graphs and Self Organizing Maps [10]. In the present work the aim is placed in mimicking some of the most plausible physiological mechanisms used in the Auditory Pathways and Centres of Human Perception for vowel spotting and characterization [11]. The detection and characterization of vowel spaces is of most importance in many applications, as in pathological characterization or forensic speaker recognition, therefore the present work will concentrate in specific vowel representation space detection and characterization by neuromorphic methods. The paper is organized as follows: A brief description of vowel nature based in formant characteristics and dynamics is given in section 2. In section 3 the layers of a Neuromorphic Speech Processing Architecture

based on Hebbian Units [7] implementing the detection paradigms is presented. In section 4 some results are given from simulations, accompanied by a brief discussion. Conclusions are presented in section 5.

2 Nature and Structure of Vowels

Speech may be described as a time-running acoustic succession of events (or phonetic sequence, see Fig.2.top) [7]. Each event is associated with an oversimplified phonation paradigm composed of vowels, and non-vowels. The acoustic-phonetic nature of these beads is based on the association of the two first resonances of the Vocal Tract, which are referred to as 'formants', and described as F_1 and F_2. F_1 in the range of 200-800 Hz is the lowest,. F_2 sweeps a wider range, from 500 to 3000 Hz. Under this point of view the nature of vowels may be described by formant stability during a time interval larger than 30 ms, and relative position in the F_2 vs F_1 space, in which is often called the 'Vowel Triangle' (see Fig.1).

Non-vowel sounds are characterized by unstable formants (dynamic), by not having a representation inside the vowel triangle, or by lacking a neat F_2 vs F_1 pattern. Sounds as $[\omega, j, b, d, J, g, p, t, c, k, \beta, \delta, \zeta, \gamma, r, r]$ are included in the first class. The second class comprises vowel-like sounds by their stability as $[l, \lambda, \Upsilon, v, z, m, n, n, \eta]$ but with representation spaces out of the area delimited by the triangle [i, a, u]. The third group includes unvoiced sounds as $[f, s, \varphi, \theta, \int, \chi,]$ which are articulated without phonation (vocal fold vibration) and produce smeared pseudo-formants in the spectrum resulting from turbulent air flow in the vocal tract. The International Phonetic Alphabet (IPA) [2] has been used,

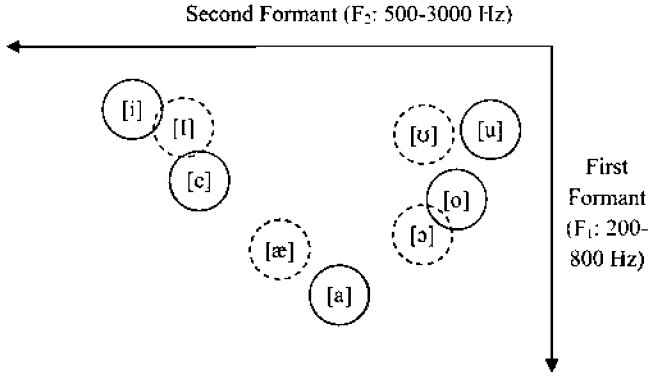

Fig. 1. Subset of the Reference Vowel Triangle for the case under study. The plot of F_2 (ordinate) vs F_2 (abscissa) is the one classically used in Linguistics. The vowel set i, e, a, o, u is sometimes referred as the *cardinal set*. The number of vowels differentiated by a listener (full line) depends on the phonologic coding of each language. Other acoustic realizations (dash line) are commonly assigned to nearby phonologic representations. For instance, in the case of study the acoustic realization [æ] in Spanish could be perceptually assigned by a listener to /a/.

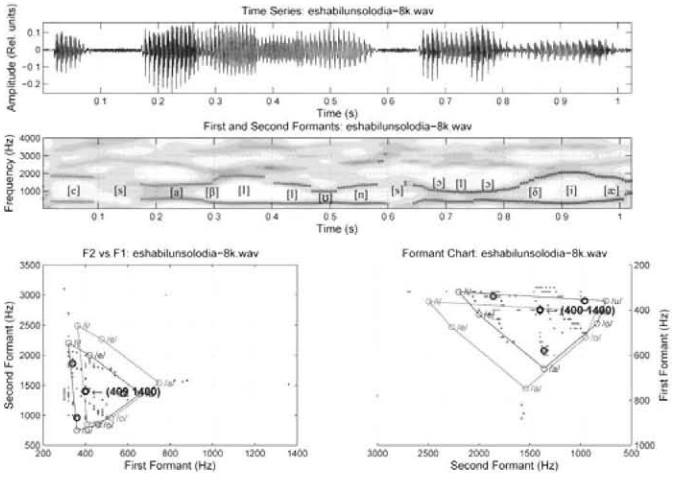

Fig. 2. Top: time series of the utterance -*es hábil un solo día*- ([*esaβIωvnsolodiæ*]) uttered by a male speaker. Middle: Adaptive Linear Prediction Spectrogram (grey background) and first two formants (superimposed in color). The color dots mark the positions of each pair (F_1,F_2) from green (the oldest) to red (the most recent). An approximate phonetic labeling is given as a reference. Bottom Left: Formant plot of F_2 vs F_1. Bottom Right: Same plot as a Formant Chart commonly used in Linguistics. The black circles give the centroids of the vowel triangle extremes and its center of gravity. The blue triangle and circles give the limit positions of the five cardinal vowels /i/, /e/, /a/, /o/, /u/ (male speaker in blue, female in magenta). These plots show the formant trajectories of the utterance. There is color correspondence between the bottom and middle templates to track formant trajectories on the time axis.

with symbols between square brackets [a] and bars /a/ are phonemes (acoustic representations) and phonologic representations, respectively. A target sentence is used as an example in Fig.2 which reproduces a spectrogram with both static and dynamic formant patterns. The sentence -*es hábil un solo día*- represents the full vowel triangle in Spanish, although acoustically some of the vowels are not extreme. Formants are characterized in this spectrogram (middle template) by darker energy envelope peaks. What can be observed in the figure is that the vowels and vowel-like sounds correspond to stable positions of the formants.

3 Neuromorphic Computing for Speech Processing

The term 'neuromorphic' is used for emulating information processing by neurologic systems. As far as speech is concerned, it has to see with neuronal units and circuits found in the Auditory Pathways and Centres. The functionality of these structures is becoming better understood as neurophysiology is deepening in functionality [3][13][15]. Preliminary work has been carried out on the

characterization of speech dynamics by the Auditory Cortex for consonant description [4][5], where a Neuromorphic Speech Processing Architecture (NSPA) based in Hebbian Units [7] was proposed and widely discussed. The present paper is focussed on the sections of the NSPA specifically devoted to vowel characterization. A general description is given in Fig.3.

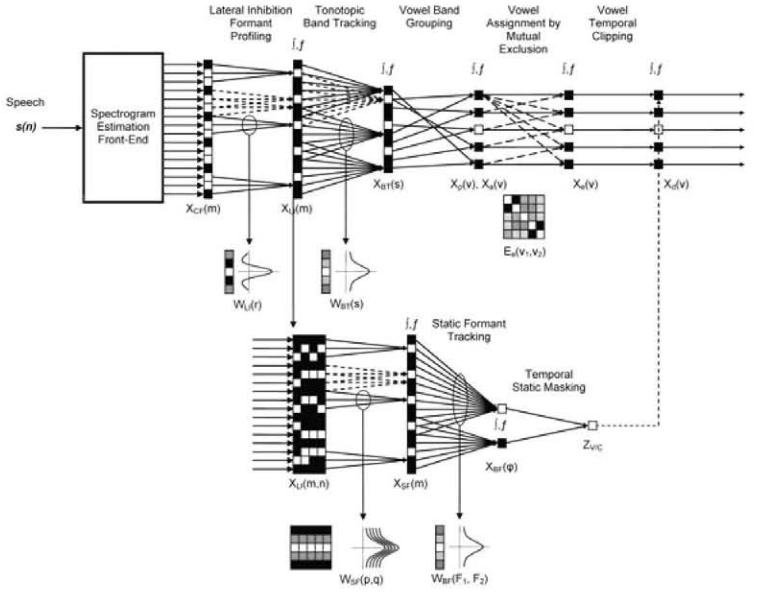

Fig. 3. Vowel processing and representation sections of the Neuromorphic Speech Processing Architecture described in [4][5]. Upper data-flow pipeline: Spectrogram Estimation Front-End, Lateral Inhibition Formant Profiling, Tonotopic Band Tracking, Vowel Band Grouping, Vowel Assignment by Mutual Exclusion and Vowel Temporal Clipping. Lower data-flow pipeline: Static Formant Tracking and Temporal Static Masking (see text for a detailed description).

Spectrogram Estimation Front-End. This section provides a spectral description of speech s(n) evolving in the time domain (spectrogram as the one in Fig.2.middle). A matrix $X_{CF}(m,n)$ is produced describing frequency activity in time (where n is the time index) as a result of a linear layer of characteristic frequency (CF) units. These units may be seen as roughly related to nerve fibres in the Auditory Periphery, each one reacting to a specific channel in frequency (where m is the frequency index). In the present case Linear Predictive Coding have been used to build the spectrogram:

$$X_{CF}(m,n) = 20 \cdot log_{10} \left| 1 - \sum_{k=1}^{K} a_{k,n} e^{-jmk\Omega\tau} \right|^{-1} \quad (1)$$

where $a_{k,n}, 1 \leq k \leq K$ is the set of coefficients of the equivalent *K-order* Inverse Filter, Ω the frequency resolution (separation between channels) and τ the sampling interval.

Lateral Inhibition Formant Profiling. The activity of neighbour fibres is reduced to represent formant descriptions at the lowest cost by lateral inhibition [6] as:

$$X_{LI}(m) = u\left(\sum_{i=-r}^{r} w_{LI}(i) X_{CF}(m+i) - \vartheta_{LI}(m)\right) \tag{2}$$

where w_i are the weights in the lateral inhibition connections. Typically, for a set of five weights (r=2) these may be set up to configurations such as $-1/6, -1/3, 1, -1/3, -1/6$, reproducing the classical Mexican Hat. The function implicit in (2) may be seen as a Hebbian Unit modelling membrane integration and threshold ($\int$,) by weighted average and nonlinear conforming. Therefore *u(.)* is a nonlinear activation function (step or sigmoid) firing if membrane activity overcomes a specific threshold $\vartheta_{LI}(m)$.

Tonotopic Band Tracking. Vowel detection is based on the combination of activity by band tracking units (BTU's) from neighbour CF fibers by Hebbian Units as:

$$X_{BT}(s) = u\left(\sum_{i=-\beta_s}^{\beta} w_{BT}(i,s) X_{LI}(\gamma_s + i) - \vartheta_{LI}(s)\right) \tag{3}$$

where s is the band index, γ_s and β_s are the indices to the center frequency and half the bandwidth respectively. In this case, the weights of the summation w_{BT} are selected to reproduce the output probability of the band according to a marginal probability density function (gaussian, with μ_s and σ_s the band mean and standard deviation):

$$\begin{aligned} &X_{BT}(i,s)) = \Gamma(\xi_i \mid \mu_s, \sigma_s) = \frac{1}{\sigma_s\sqrt{2\pi}} e^{-\frac{(\xi_i - \mu_s)^2}{2\sigma_s^2}} \\ &-\beta_s\Omega \leq \xi_i \leq -\beta_s\Omega; \xi_i = i\Omega; \mu_s = \gamma_s\Omega; \sigma_s = \beta_s\Omega \end{aligned} \tag{4}$$

Vowel Band Grouping. Once a sufficient number of BTU's have tuned their respective frequency spaces, they must be somehow combined among themselves to represent vowel activity as ordered pairs $X_{BT}(i), X_{BT}(j)$. This combination strategy is very much language-dependent, based on a previous agreement among the speakers of the language. As a matter of fact each language has developed its own encoding table, which finds its counterpart in the representation spaces to be found in the Auditory Centers. As an example, the encoding table for the five cardinal vowels [a, e, i, o, u] for standard Spanish is shown in Table 1. Other languages are known to have a larger symbol system, in which case the phonological vowel set would be correspondingly larger.

Table 1. Phonol. Formant Association Table for Spanish

BTU's F_2/F_1 (Hz)	**220-440**	**300-600**	**550-950**
550-850	/u/	void	void
700-1100	aliased	/o/	void
900-1500	aliased	aliased	/a/
1400-2400	aliased	/e/	void
1700-2900	/i/	void	void

This configuration is the result of averaging estimations from 8 male speakers, a similar table for female speakers could be produced. The positions marked as 'void' correspond to non-vowel sounds (second class), whereas the positions marked as 'aliased' may be ascribed to nearby valid vowel representation spaces showing a larger probability function with respect to the acoustic model.

Vowel Assignment by Mutual Exclusion. The vowel representation spaces must be unambiguously coded to bear plausible meaning to the listener. Therefore a strong exclusion mechanism is proposed, which would be activated each time enough activity is detected simultaneously by several units in a specific acoustic space, thus the vowel showing the largest activity or detection probability reacts as a 'winner-takes-all' silencing other possible vowel candidates. A neural circuit combines each two band activities by pairs according to the following paradigm:

$$\begin{aligned} X_p(\nu) &= w_{p1}(X_{BT}(s_1) \times w_{p2}(X_{BT}(s_2) \\ X_a(\nu))) &= u(X_p(\nu) - \vartheta_a(\nu)) \end{aligned} \tag{5}$$

where $X_p(\nu)$ may be seen as the activation probability for vowel v given the input template $X_{CF}(m)$, and ν is the index to the set of vowels in the phonological system:

$$X_p(\nu) = p(\nu \mid X_{CF}(m)); \nu \in \{u, o, a, i, e\} \tag{6}$$

On its turn, weights w_{p1} and w_{p2} encode the relative probabilities of the respective formants in the detection of the vowel. The symbol $(\times)$ represents the logical operator and, and may be implemented also by a Hebbian Unit. The mutual exclusion among representation spaces is governed by the following combination paradigm:

$$X_a(\nu))) = u(E(\nu_1, \nu_2)X_a(\nu) - \vartheta_e(\nu)) \tag{7}$$

where $E(v_1, v_2)$ is the mutual exclusion matrix, pre-wired as in the present case:

$$E(v_1, v_2) = \begin{pmatrix} +1.0 & -1.0 & +0.0 & 0.0 & -0.2 \\ -1.0 & +1.0 & -0.2 & -0.2 & +0.0 \\ +0.0 & -1.0 & +1.0 & -1.0 & +0.0 \\ +0.0 & -0.2 & -0.2 & +1.0 & -1.0 \\ -0.2 & +0.0 & +0.0 & -1.0 & +1.0 \end{pmatrix} \tag{8}$$

The elements in the main diagonal are set to +1.0, each vowel probability exciting the next unit (solid arrow in Fig.3) whereas it acts as a strong, weak or neuter inhibitory input (-1.0, -0.2, +0.0) to other vowels (dash arrows). Equation (7) is a discriminant function [8] based on Bayesian Decision Theory using log likelihood ratios:

$$L_e(\nu) = log\left\{\frac{p(x_{CF} \mid \nu}{p(x_{CF} \mid \bar{\nu}}\right\}; X_e(\nu) = \begin{cases} 1; L_e(\nu) > \xi_e(\nu) \\ 0; L_e(\nu) \leq \xi_e(\nu) \end{cases} \tag{9}$$

Vowel Temporal Clipping. This step adds the stability property demanded for vowel sounds. A control signal as $Z_{V/C}(n)$ marking the temporal segments or intervals where formants are stable within some limits is used to inhibit or enable the expression of each vowel by logical and functions ($\times$) as defined in (5):

$$X_d(\nu) = u\left(Z_{V/C} \times X_e(\nu) - \vartheta_d(\nu)\right) \tag{10}$$

Static Formant Tracking. The temporal clipping signal is estimated by tracking the segments where the first two formants remain relatively stable. This activity is captured using mask-based neuromorphic units as already explained in [4][5] which process the spectrogram as a true auditory image [8]:

$$X_{SF}(m,n) = u\left[\sum_{p=-P}^{P}\sum_{q=0}^{Q} w_{SF}(p,q)X(m+p,n-q)) - \vartheta_{SF}(m)\right] \tag{11}$$

The weight matrix $w_{SF}(p,q)$ is a bell-shaped histogram displaced in the time index (q). Practical values for P and Q are 4 and 8, respectively, resulting in a 9x9 mask.

Temporal Static Masking. Stability has to be detected separately on the two first formants and further combined. Two independent units, φ_1 and φ_1 will be tuned to two frequency bands centred at (γ_1, γ_2) with half bandwidths (β_1, β_2) similarly to (3):

$$X_{BF}(\varphi) = u\left(\sum_{i=-\beta_s}^{\beta} w_{BF}(i,\varphi)X_{SF}(\gamma_\varphi + i) - \vartheta_{BF}(\varphi)\right) \tag{12}$$

The weights of the integration function are fixed as gaussian distributions following (4). The fusion of formant masking units is carried out by a classical *and* operator:

$$Z_{V/C} = u\left(X_{BF}(\varphi_1) \times X_{BF}(\varphi_1) - \vartheta_{V/C}\right) \tag{13}$$

This signal is used in (10) to validate the intervals of formant stable activity which can be associated to vowel representation spaces.

Table 1. Phonol. Formant Association Table for Spanish

BTU's F_2/F_1 (Hz)	**220-440**	**300-600**	**550-950**
550-850	/u/	void	void
700-1100	aliased	/o/	void
900-1500	aliased	aliased	/a/
1400-2400	aliased	/e/	void
1700-2900	/i/	void	void

This configuration is the result of averaging estimations from 8 male speakers, a similar table for female speakers could be produced. The positions marked as 'void' correspond to non-vowel sounds (second class), whereas the positions marked as 'aliased' may be ascribed to nearby valid vowel representation spaces showing a larger probability function with respect to the acoustic model.

Vowel Assignment by Mutual Exclusion. The vowel representation spaces must be unambiguously coded to bear plausible meaning to the listener. Therefore a strong exclusion mechanism is proposed, which would be activated each time enough activity is detected simultaneously by several units in a specific acoustic space, thus the vowel showing the largest activity or detection probability reacts as a 'winner-takes-all' silencing other possible vowel candidates. A neural circuit combines each two band activities by pairs according to the following paradigm:

$$\begin{aligned} X_p(\nu) &= w_{p1}(X_{BT}(s_1) \times w_{p2}(X_{BT}(s_2) \\ X_a(\nu))) &= u(X_p(\nu) - \vartheta_a(\nu)) \end{aligned} \tag{5}$$

where $X_p(\nu)$ may be seen as the activation probability for vowel v given the input template $X_{CF}(m)$, and ν is the index to the set of vowels in the phonological system:

$$X_p(\nu) = p(\nu \mid X_{CF}(m)); \nu \in \{u, o, a, i, e\} \tag{6}$$

On its turn, weights w_{p1} and w_{p2} encode the relative probabilities of the respective formants in the detection of the vowel. The symbol $(\times)$ represents the logical operator and, and may be implemented also by a Hebbian Unit. The mutual exclusion among representation spaces is governed by the following combination paradigm:

$$X_a(\nu))) = u(E(\nu_1, \nu_2)X_a(\nu) - \vartheta_e(\nu)) \tag{7}$$

where $E(v_1, v_2)$ is the mutual exclusion matrix, pre-wired as in the present case:

$$E(v_1, v_2) = \begin{pmatrix} +1.0 & -1.0 & +0.0 & 0.0 & -0.2 \\ -1.0 & +1.0 & -0.2 & -0.2 & +0.0 \\ +0.0 & -1.0 & +1.0 & -1.0 & +0.0 \\ +0.0 & -0.2 & -0.2 & +1.0 & -1.0 \\ -0.2 & +0.0 & +0.0 & -1.0 & +1.0 \end{pmatrix} \tag{8}$$

4 Results and Discussion

In what follows some results from processing the model sentence in Fig.2 with the proposed structure will be shown. The details of the architecture are the following: $1 \leq m \leq M = 512$ CF fibre units are used, defining a resolution in frequency of 16 Hz for a sampling frequency of 8000 Hz. A spectrum frame is produced each 2 ms to define a stream of approximately 500 frames per second. The dimensions of the BTU's are defined as in Table 1. An example of the operation of BTU's $X_{BT}(220-440)$ and $X_{BT}(1800-3000)$ and the formant fusion unit $X_a(/i/)$ is shown in Fig.4.

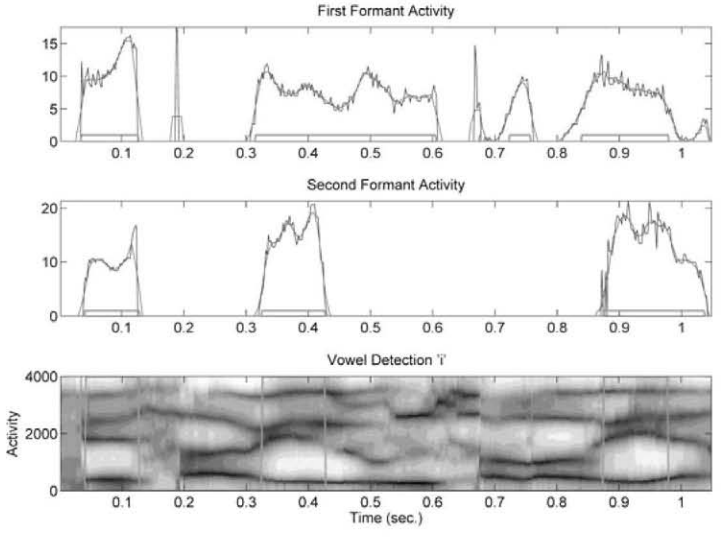

Fig. 4. Top: Activity of BTU $X_{BT}(220-440)$. Input activity at the unit membrane before (blue) and after integration (red), and firing after threshold (green). Middle: Idem for $X_{BT}(1800-3000)$. Bottom: Fusion of both BTU's in unit $X_a(/i/)$ (in green). The spectrogram is given as a reference.

This unit selects vowel segments corresponding to [I] or [i], and to [e] (first segment between 0.04 and 0.13 s). This is compliant with the ability of any BTU to capture activity from acoustic spaces overlapping in part with neighbour units as explained before. When the respective activities of both $X_a(/e/)$ and $X_a(/i/)$ are subject to mutual exclusion the first segment will be assigned to /e/ (cyan) and the two last ones will be captured by /i/ (blue) as seen in Fig.5. Vowel detection is evident after this operation.

The use of the temporal static masking signal $Z_{V/C}$ helps in removing certain ambiguities in vowel-consonant assignments as it may be seen in Fig.6. The vowel intervals have been delimited to the most stable segments of the utterance. Table 2 gives a detailed description of the detection process.

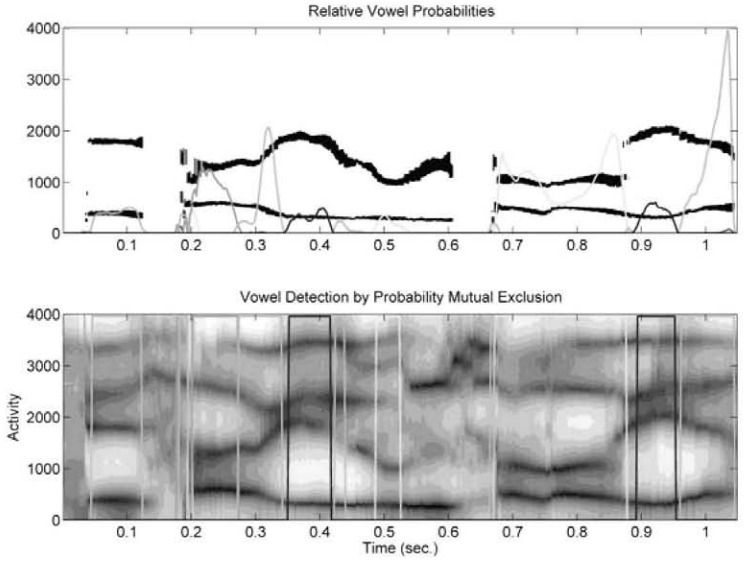

Fig. 5. Top: Probability estimates for the five vowels at layer $X_a(\nu)$. The first two formants are superimposed for reference as by layer $X_{LI}(m)$. Bottom: Activity of layer $X_e(\nu)$. Vowel color reference: /i/-blue, /e/-cyan, /a/-green, /o/-yellow, /u/-red.

Table 2. Vowel detection results

Interval (ms)	Observations
0.04-0.13	[e] is detected
0.13-0.21	void (sibilant [s])
0.21-0.27	[a] is detected
0.27-0.30	[æ] is detected as /e/
0.31-0.35	void (approximant [β])
0.35-0.41	[i] is detected
0.41-0.50	void (lateral [l])
0.50-0.53	[υ] is detected as /o/
0.53-0.69	void (nasal [n] and a sibilant [s])
0.69-0.76	[o] is detected as /o/
0.76-0.77	void (lateral [l])
0.77-0.86	[o] is detected as /o/
0.86-0.89	void (approximant [δ])
0.89-0.96	[i] is detected
0.96-1.05	unstable [$i \rightarrow e \rightarrow$æ] is fragmentarily detected as /e/

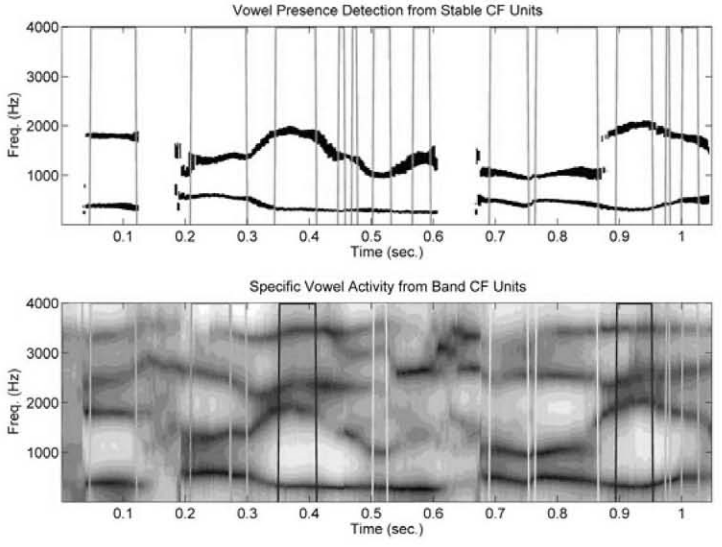

Fig. 6. Top: Output activity of the temporal masking unit $Z_{V/C}$. Bottom: Activity of layer $X_d(\nu)$

5 Conclusions

Through the present work it has been shown that vowel characterization can be carried out based on the criteria of formant stability and relative position inside the vowel triangle of the speaker using neuromorphic (Hebbian) processing units (neurons). It has also been shown that band categorization is carried out using gaussians as marginal distributions. Under this point of view the membrane activity of band categorization neurons (after integration) may receive the consideration of conditional probabilities. Output firing rates are to be seen as results of decision-making algorithms when mutual exclusion is used on competing conditional probabilities. The process relies strongly on the use of lateral inhibition to profile formants and to establish vowel representation spaces in a "winner-takes-all" strategy. This implies a decision problem which may produce unexpected results, as in the interval 0.50-0.53, where a rather obscure vowel [ʊ] is mistaken as /o/. This fact demands a small explanation: although the resulting vowel space is not fully represented by /o/ the acoustic-phonetic space controlled by this symbol is very ubiquitous, as to be able of seizing the surrounding space, which is not very much questioned by any of the other vowel representations except /u/, -see the mutual exclusion matrix in (8). This result is left deliberately 'as-is' to put into evidence eager seizing (aliasing or usurpation) of unclaimed representation spaces by strongly implanted vowels under the phonological point of view. This behaviour may explain difficulties in speakers of reduced vowel representation spaces to recognize much richer vowel systems from foreign origin. The utility of these results is to be found in automatic

phonetic labeling of the speech trace for speech spotting, as well as in the detection of the speaker's identity [14], where stable characteristic vowel segments are sought for contrastive similarity tests.

Acknowledgements

This work is being funded by grants TEC2009-14123-C04-03 from Plan Nacional de I+D+i, Ministry of Science and Technology of Spain and CCG06-UPM/TIC-0028 from CAM/UPM.

References

1. Acero, A.: New Machine Learning Approaches to Speech Recognition. In: FALA 2010, Vigo, Spain, November 10-12 (2010); ISBN: 978-84-8158-510-0
2. http://www.arts.gla.ac.uk/IPA/ipachart.html
3. Barbour, D.L., Wang, X.: Temporal Coherence Sensitivity in Auditory Cortex. J. Neurophysiol. 88, 2684–2699 (2002)
4. Gómez, P., Ferrández, J.M., Rodellar, V., Fernández, R.: Time-frequency Representations in Speech Perception. Neurocomputing 72, 820–830 (2009)
5. Gómez, P., Ferrández, J.M., Rodellar, V., Alvarez, A., Mazaira, L.M., Olalla, R., Muñoz, C.: Neuromorphic detection of speech dynamics. Neurocomputing 74(8), 1191–1202 (2011)
6. Greenberg, S., Ainsworth, W.H.: Speech processing in the auditory system: an overview. In: Greenberg, W.A.S. (ed.) Speech Processing in the Auditory System, pp. 1–62. Springer, New York (2004)
7. Hebb, D.O.: The Organization of Behavior. Wiley, New York (1949)
8. Huang, X., Acero, A., Hon, H.W.: Spoken Language Processing. Prentice-Hall, Upper Saddle River (2001)
9. Jahne, B.: Digital Image Processing. Springer, Berlin (2005)
10. Kohonen, T.: Self-Organizing Maps. Springer, Heidelberg (1997)
11. Munkong, R., Juang, B.H.: Auditory Perception and Cognition. IEEE Signal Proc. Magazine, 98–117 (May 2008)
12. O'Shaughnessy, D.: Speech Communication. Human and Machine. Addison-Wesley, Reading (2000)
13. Palmer, A., Shamma, S.: Physiological Representation of Speech. In: Greenberg, S., Ainsworth, W., Popper, A. (eds.), pp. 163–230. Springer, New York (2004)
14. Rose, P., Kinoshita, Y., Alderman, T.: Realistic Extrinsic Forensic Speaker Discrimination with the Diphthong /aI/. In: Proc. 11th Austr. Int. Conf. on Speech Sci. and Tech., pp. 329–334 (December 2006)
15. Shamma, S.: Physiological foundations of temporal integration in the perception of speech. J. Phonetics 31, 495–501 (2003)